Replication Variance Estimation for Imputed Data

A. Chatterjee[†], W.A. Fuller and J.D. Opsomer
Center for Survey Statistics and Methodology
and
Department of Statistics, Iowa State University, Ames, IA 50011

Abstract

Replication methods are used frequently to construct variance estimators under complex sampling designs. We develop a replication variance estimator for the situation in which some of the sample elements are unobserved and their information is imputed. The method relies on fractional hot-deck imputation and adjusts the variance replicates to account for the imputation mechanism. The properties of the method are discussed.

Keywords: Hot deck imputation, fractional imputation, jackknife variance estimation.

1 Introduction

Item nonresponse occurs frequently in surveys. In many situations, hot deck imputation procedures are used to impute the missing values. These procedures replace the missing values by values randomly selected from respondents. In that case, a full dataset without missing values is obtained, with some values actually observed and other imputed. If variance estimates using standard formulas are then used for estimating variances, it can lead to underestimation of the variability of the survey estimates. An approach for reducing imputation variance is to use *fractional imputation* as described by Kalton and Kish (1984), Fay (1996), and Kim and Fuller (2004). Fractional imputation involves using more than one donor for a nonrespondent, with each donor having fractional weights. This approach was designed to reduce the variance of the imputation procedure, but it also provides an opportunity to estimate the variability of the procedure by considering the between-imputation differences, as will be further explored below.

Several other procedures have been proposed for estimating the variance of an estimated total after hot deck imputation. Rao and Shao (1992) proposed a jackknife variance estimator for hot deck imputation in which the donors are selected with probability proportional to the sampling weights. Tollefson and Fuller (1992) proposed a variance estimator for without replacement hot deck imputation. Särndal (1992), Fay (1996) and Chen and Shao (2001) also proposed variance estimators for certain types of hot deck imputation. For more references, see Little and Rubin (2002).

The impetus for the current research come from the involvement of the authors in the development of variance estimation procedures for the National Resources Inventory (NRI). The NRI is a two-phase longitudinal survey designed to assess conditions and trends of soil, water and related natural resources on nonfederal lands in the United States. The NRI is being conducted by the US Department of Agriculture's Natural Resources Conservation Service, in cooperation with the Center for Survey Statistics and Methodology at Iowa State University. Nusser and Goebel (1997) describe the NRI in more detail. Originally conducted every five years, the current NRI is conducted annually, with the 1997 NRI used as a first phase sample and the annual samples selected as second phase samples. While a portion of the sampled elements have complete data records for all survey years, another portion has incomplete records over time. For the latter portion, a hierarchical hot deck procedure is used to impute the missing data. Regression and calibration estimation procedures are used to construct annual estimates, and a grouped jackknife procedure is used to estimate the variance. This variance estimation procedure does not currently fully account for the variance inflation due to imputation, so there is a need for an improved procedure that can account for that effect.

The estimation procedure described in the following sections proposes an "idealized" version of a procedure that might eventually be adapted for the NRI. In this simplified version, the sampling design is simple random sampling without replacement, the response mechanism is simple random sampling without replacement within groups, and the imputation mechanism is through equal-probability with-replacement sampling within the same groups. These simplifications were chosen to make it easier to explain the approach and to formally derive its

[†]204 Snedecor Hall, Iowa State University, Ames, IA 50011; cha@iastate.edu.

statistical properties. In the future, we intend to expand the study of the proposed estimation procedure to account for more complex sampling designs and imputation mechanisms.

In Section 2 we describe the fractional imputation based estimator. In Section 3 we discuss the properties of this estimator and related results. In Section 4, we propose an adjusted variance estimator and in Section 5 we show the results of a small simulation study and finally in Section 6, we present our conclusions.

2 Fractional Imputation Estimator

We assume that a without replacement sample of size n is drawn from a (large) population U containing N elements. In this sample there are n_R respondents and n_M missing values. For the purpose of imputing the missing values, assume that we are constructing G imputation cells, in such a way that the observations are approximately homogenous within each cell. We denote the sample of size n as s. We assume that these G imputation groups extend into the population U and in the gth group there are N_g elements, with $\sum_{g=1}^{G} N_g = N$. The sample s of size n has n_g elements drawn from the gth group for all $g = 1, \ldots, G$. Hence the n_g's are random and $\sum_{g=1}^{G} n_g = n$.

Among the n_g elements sampled out of the gth group, there are n_{R_g} respondents and n_{M_g} missing values with, $n_{R_g} + n_{M_g} = n_g$. We assume that these n_{R_g} respondents are drawn without replacement from the n_g elements with the response set in any group being independent of any other group. We denote the set of responding units (sample) from the gth group as R_g. The combined set of respondents over all groups is denoted as

$$R = \cup_{g=1}^{G} R_g.$$

Also, call the full set of missing values as M, with M_g the set of missing values in group g.

We will maintain a design based framework and study the proposed imputation method with respect to the (random) imputation mechanism, the response mechanism and the sampling design. The imputation mechanism we will consider in this paper is *with-replacement partial fractional hot deck imputation within cells*.

Under fractional imputation, the missing values are replaced by several randomly imputed values with fractional weights. In partial fractional imputation, some missing values will be fractionally imputed while others will only be imputed once. We divide the elements in the class M_g into two classes denoted by $M_g^{(1)}$ and $M_g^{(2)}$, where the values in the former class are imputed by using only one randomly chosen obsevation from R_g and those in the latter class are imputed by selecting two observations with replacement from R_g. The sample sizes in these classes are denoted by $n_{M_g}^{(1)}$ and $n_{M_g}^{(2)}$, respectively, and we denote their totals over the cells as $n_M^{(1)} = \sum_{g=1}^{G} n_{M_g}^{(1)}$ and $n_M^{(2)} = \sum_{g=1}^{G} n_{M_g}^{(2)}$. Also, $n_M = n_M^{(1)} + n_M^{(2)} = n - n_R = \sum_g n_g - \sum_g n_{R_g}$, and, $n_{R_g} + n_{M_g}^{(1)} + n_{M_g}^{(2)} = n_g$. We assume that, after the respondents R_g are selected, $n_{M_g}^{(1)}$ units are randomly selected from the set of n_{M_g} nonrespondents. Thus, the set of missing values which are imputed using only one observation is a random sample (of size $n_{M_g}^{(1)}$) from the full set of nonresponding units. The other set of $n_{M_g}^{(2)}$ elements which are imputed using two observations is automatically chosen.

To further simplify the notation, we assume that the following proportionality conditions hold over all groups g,

$$n_{R_g} = \alpha_1 n_g \quad \text{and} \quad n_{M_g}^{(1)} = \alpha_2 n_g \ \ \forall\ g \qquad (1)$$

for some constants $\alpha_1, \alpha_2 \in (0, 1)$ with $(\alpha_1 + \alpha_2) < 1$. This condition means that the proportion of respondents and units imputed with one (or two) respondents remain the same over all groups. This assumption can be readily relaxed, but would make our notation significantly more cumbersome.

We consider the estimation of

$$\bar{y}_N = \frac{1}{N} \sum_{i=1}^{N} y_i \equiv \frac{1}{N} \sum_{g=1}^{G} \sum_{j=1}^{N_g} y_{gj}.$$

The estimator we will study is,

$$\bar{y}_I = \frac{1}{n} \sum_{g} \sum_{j} y_{gj}^{**}. \qquad (2)$$

where

$$y_{gj}^{**} = \begin{cases} y_{gj} & : \ j \in R_g \\ y_{gj}^{*} & : \ j \in M_g^{(1)} \\ \frac{1}{2}(y_{gj1}^{*} + y_{gj2}^{*}) & : \ j \in M_g^{(2)} \end{cases}$$

and the values $y_{gj}^{*}, y_{gj1}^{*}, y_{gj2}^{*}$ are drawn with replacement and with equal probability from those in R_g.

3 Properties of the Estimator

We study the properties of the estimator (2) with respect to the sampling design, the response mechanism and imputation mechanism. Let $\mathrm{E}_{im}\left(.\middle|R\right)$ denote the expectation with respect to the imputation

mechanism given the response set R and $\mathrm{V}_{im}\left(.\middle|R\right)$ denotes the corresponding variance (conditioning on the response set R trivially conditions on the sample s as well). Similarly, $\mathrm{E}_R\left(.\middle|s\right)$ and $\mathrm{V}_R\left(.\middle|s\right)$ denote the expectation and variance with respect to the 'second phase' sample R given the first phase sample s. $\mathrm{E}_s(.)$ and $\mathrm{V}_s(.)$ denote expectation and variance with respect to the first phase sampling. Finally, E(.) and V(.) denote overall expectation and variance. We can write

$$\mathrm{E}(\bar{y}_I) = \mathrm{E}_s\left(\mathrm{E}_R\left(\mathrm{E}_{im}\left(\bar{y}_I\middle|R\right)\middle|s\right)\right),$$

and

$$\begin{aligned}\mathrm{V}(\bar{y}_I) &= \mathrm{V}_s(\mathrm{E}_R(\bar{y}_I\big|s)) + \mathrm{E}_s(\mathrm{V}_R(\bar{y}_I\big|s)) \\ &= \mathrm{V}_s(\mathrm{E}_R(\mathrm{E}_{im}(\bar{y}_I\big|R)\big|s)) \\ &\quad + \mathrm{E}_s\left[\mathrm{V}_R(\mathrm{E}_{im}(\bar{y}_I\big|R)\big|s)\right. \\ &\quad \left. + \mathrm{E}_R(\mathrm{V}_{im}(\bar{y}_I\big|R)\big|s)\right]. \qquad (3)\end{aligned}$$

Conditioning on R,

$$\mathrm{E}_{im}\left(y^{**}_{gj}\middle|R\right) = \begin{cases} y_{gj} & : \quad j \in R_g \\ \bar{y}_{R_g} & : \quad j \in M_g \end{cases}$$

and

$$\mathrm{V}_{im}\left(y^{**}_{gj}\middle|R\right) = \begin{cases} 0 & : \quad j \in R_g \\ s^2_{R_g} & : \quad j \in M_g \end{cases},$$

with $\bar{y}_{R_g} = \frac{1}{n_{R_g}}\sum_{j\in R_g} y_{gj}$ and $s^2_{R_g} = \frac{1}{n_{R_g}-1}\sum_{j\in R_g}\left(y_{gj}-\bar{y}_{R_g}\right)^2$. Since the imputation is done with replacement, the covariance between any two imputed values is zero, so that under the assumptions made in (1),

$$\begin{aligned}\mathrm{E}_R(\mathrm{E}_{im}(\bar{y}_I\big|R)\big|s) &= \mathrm{E}_R\left(\frac{1}{n}\sum_g n_g\bar{y}_{R_g}\middle|s\right) \\ &= \frac{1}{n}\sum_g n_g\bar{y}_{n_g}, \qquad (4)\end{aligned}$$

$$\begin{aligned}\mathrm{V}_R(\mathrm{E}_{im}(\bar{y}_I\big|R)\big|s) &= \mathrm{V}_R\left(\frac{1}{n}\sum_g n_g\bar{y}_{R_g}\middle|s\right) \\ &= \frac{1}{n^2}\sum_g n_g^2\left(1-\frac{n_{R_g}}{n_g}\right)\frac{s^2_{n_g}}{n_{R_g}} \\ &= \frac{1}{n}\frac{n_M}{n_R}\sum_g \frac{n_g}{n}s^2_{n_g} \qquad (5)\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}_R(\mathrm{V}_{im}(\bar{y}_I\big|R)\big|s) &= \frac{1}{n^2}\sum_g \tilde{n}_{M_g}s^2_{n_g} \\ &= \frac{\tilde{n}_M}{n^2}\sum_g \frac{n_g}{n}s^2_{n_g}. \qquad (6)\end{aligned}$$

where $s^2_{n_g} = \frac{1}{n_g-1}\sum_{j=1}^{n_g}\left(y^{**}_{gj}-\bar{y}_{n_g}\right)^2$ and

$$\tilde{n}_{M_g} = \left(n^{(1)}_{M_g} + \frac{n^{(2)}_{M_g}}{2}\right).$$

Since s was a without replacement sample, we can deduce from (3) and (4),

$$\mathrm{V}_s(\mathrm{E}_R(\mathrm{E}_{im}(\bar{y}_I\big|R)\big|s)) = \mathrm{V}_s(\bar{y}_n) = \left(1-\frac{n}{N}\right)\frac{s^2_N}{n}$$

where $s^2_N = \frac{1}{N-1}\sum_{g=1}^{G}\sum_{j=1}^{N_g}\left(y_{gj}-\bar{y}_N\right)^2$ and $\bar{y}_N = \frac{1}{N}\sum_{g=1}^{G}\sum_{j=1}^{N_g} y_{gj}$. Combining (5) and (6) we get

$$\mathrm{E}_s(\mathrm{V}_R(\bar{y}_I\big|s)) = \left(\frac{\tilde{n}_M}{n^2}+\frac{1}{n}\frac{n_M}{n_R}\right)\mathrm{E}_s\left(\sum_g \frac{n_g}{n}s^2_{n_g}\right).$$

Hence, (3) reduces to

$$\begin{aligned}\mathrm{V}(\bar{y}_I) &= \left(1-\frac{n}{N}\right)\frac{s^2_N}{n} + \left(\frac{n_M}{n^2} - \frac{n^{(2)}_M}{2n^2} + \frac{1}{n_R} - \frac{1}{n}\right) \\ &\quad \times \mathrm{E}_s\left(\sum_g \frac{n_g}{n}s^2_{n_g}\right). \qquad (7)\end{aligned}$$

Under the first sampling design s, the quantities n_g are random and so we keep the terms $\frac{n_g}{n}$ inside the expectation. The term $\mathrm{E}_s\left(\sum_g \frac{n_g}{n}s^2_{n_g}\right)$ is a weighted average of the within group variances. We would like to develop an estimator for the variance in (7). A possible (naive) estimator can be defined as,

$$\widehat{\mathrm{V}}(\bar{y}_I) = \frac{1}{n(n-1)}\sum_g\sum_j\left(y^{**}_{gj}-\bar{y}_I\right)^2. \qquad (8)$$

To find the expectation of this estimator we write

$$\begin{aligned}y^{**}_{gj}-\bar{y}_I &= \left(\bar{y}_{R_g} - \frac{\sum_g n_g\bar{y}_{R_g}}{n}\right) \\ &\quad + \left\{y^{**}_{gj} - \bar{y}_{R_g} - \left(\bar{y}_I - \frac{\sum_g n_g\bar{y}_{R_g}}{n}\right)\right\},\end{aligned}$$

and

$$\mathrm{E}_{im}\left(\sum_g\sum_j\left(y^{**}_{gj}-\bar{y}_I\right)^2\middle|R\right) = A + B \qquad (9)$$

where

$$\begin{aligned}A &= \sum_g n_g\left(\bar{y}_{R_g} - \frac{\sum_h n_h\bar{y}_{R_h}}{n}\right)^2 \\ &= \sum_g n_g\bar{y}^2_{R_g} - \frac{1}{n}\left(\sum_g n_g\bar{y}_{R_g}\right)^2 \qquad (10)\end{aligned}$$

and

$$
\begin{aligned}
B = & \mathrm{E}_{im}\left(\sum_g\sum_j \left(y_{gj}^{**}-\bar{y}_{R_g}\right)^2\big|R\right) \\
& + \mathrm{E}_{im}\left(n\left(\bar{y}_I-\frac{\sum_g n_g\bar{y}_{R_g}}{n}\right)^2\big|R\right) \\
& - \mathrm{E}_{im}\left(2\left(\bar{y}_I-\frac{\sum_g n_g\bar{y}_{R_g}}{n}\right)\right. \\
& \left.\times\sum_g\sum_j(y_{gj}^{**}-\bar{y}_{R_g})\big|R\right) \\
= & \ \mathrm{E}_{im}(B_1+B_2-B_3\big|R). \qquad (11)
\end{aligned}
$$

In the term B_3, the sum is actually over the units in $j \in M_g$. Taking expectations of these three terms separately and using the fact that the imputation is done with replacement, we find from B_1,

$$
\begin{aligned}
\mathrm{E}_{im}(B_1\big|R) &= \mathrm{E}_{im}\left(\sum_g\sum_j \left(y_{gj}^{**}-\bar{y}_{R_g}\right)^2\big|R\right) \\
&= \sum_g\sum_{j\in R_g}\left(y_{gj}-\bar{y}_{R_g}\right)^2 \\
&\quad + \sum_g\sum_{j\in M_g}\mathrm{E}_{im}\left(\left(y_{gj}^{**}-\bar{y}_{R_g}\right)^2\big|R\right) \\
&= \sum_g n_{R_g}s_{R_g}^2 + \sum_g\sum_{j\in M_g}\mathrm{V}_{im}(y_{gj}^{**}\big|R) \\
&= \sum_g(n_{R_g}+\tilde{n}_{M_g})s_{R_g}^2. \qquad (12)
\end{aligned}
$$

And from B_2, we get

$$
\begin{aligned}
\mathrm{E}_{im}(B_2\big|R) &= n\mathrm{E}_{im}\left\{\left(\bar{y}_I-\frac{\sum_g n_g\bar{y}_{R_g}}{n}\right)^2\big|R\right\} \\
&= n\mathrm{V}_{im}\left(\bar{y}_I\big|R\right) \\
&= \frac{1}{n}\sum_g \tilde{n}_{M_g}s_{R_g}^2. \qquad (13)
\end{aligned}
$$

Note that in B_3,

$$
\bar{y}_I-\frac{1}{n}\sum_g n_g\bar{y}_{R_g} = \frac{1}{n}\sum_g\sum_{j\in M_g}(y_{gj}^{**}-\bar{y}_{R_g}).
$$

Hence, B_3 reduces to $\frac{1}{n}\left(\sum_g\sum_{j\in M_g}(y_{gj}^{**}-\bar{y}_{R_g})\right)^2$, and

$$
\begin{aligned}
&\mathrm{E}_{im}(B_3\big|R) = \\
&\quad \mathrm{E}_{im}\left(\frac{2}{n}\left(\sum_g\sum_{j\in M_g}(y_{gj}^{**}-\bar{y}_{R_g})\right)^2\big|R\right) \\
&= \frac{2}{n}\mathrm{V}_{im}\left(\sum_g\sum_{j\in M_g}y_{gj}^{**}\big|R\right) \\
&= \frac{2}{n}\sum_g\tilde{n}_{M_g}s_{R_g}^2. \qquad (14)
\end{aligned}
$$

Hence, from (10)-(14) we get

$$
\begin{aligned}
&\mathrm{E}_{im}\left(\sum_g\sum_j(y_{gj}^{**}-\bar{y}_I)^2\big|R\right) = \sum_g n_g\bar{y}_{R_g}^2 \\
&\quad -\frac{1}{n}\left(\sum_g n_g\bar{y}_{R_g}\right)^2 + \sum_g\left(n_{R_g}+\tilde{n}_{M_g}-\frac{\tilde{n}_{M_g}}{n}\right)s_{R_g}^2. \qquad (15)
\end{aligned}
$$

The next stage is to find the conditional expectation with respect to the second phase sample R given the first phase sample s. Thus, from (8) and (9), taking expectation of (15),

$$
\begin{aligned}
&\mathrm{E}_R(\sum_g\sum_j(y_{gj}^{**}-\bar{y}_I)^2|s) = \sum_g n_g\left(1-\frac{n_{R_g}}{n_g}\right)\frac{s_{n_g}^2}{n_{R_g}} \\
&\quad + \sum_g n_g\bar{y}_{n_g}^2 - \frac{1}{n}\left[\sum_g n_g^2\left(1-\frac{n_{R_g}}{n_g}\right)\frac{s_{n_g}^2}{n_{R_g}} + n^2\bar{y}_n^2\right] \\
&\quad + \sum_g\left(n_{R_g}+\tilde{n}_{M_g}-\frac{\tilde{n}_{M_g}}{n}\right)s_{n_g}^2.
\end{aligned}
$$

Under the assumptions in (1), the right side of the

above expression reduces to

$$\begin{aligned}
\mathrm{E}_R(\sum_g \sum_j (y^{**}_{gj} - \bar{y}_I)^2 | s) &= \left(1 - \frac{n_R}{n}\right) \frac{n}{n_R} \sum_g s^2_{n_g} \\
&\quad + \sum_g n_g \bar{y}^2_{n_g} - \left(\frac{1}{n_R} - \frac{1}{n}\right) \sum_g n_g s^2_{n_g} - n\bar{y}^2_n \\
&\quad + \sum_g \left(n_{R_g} + \tilde{n}_{M_g} - \frac{\tilde{n}_{M_g}}{n}\right) s^2_{n_g} \\
&= \frac{n_M}{n_R} \sum_g \left(1 - \frac{n_g}{n}\right) s^2_{n_g} - \sum_g n_g s^2_{n_g} + n s^2_n \\
&\quad + \left(1 - \frac{n^{(2)}_M}{2n}\right) \sum_g n_g s^2_{n_g} - \frac{1}{n} \sum_g \tilde{n}_{M_g} s^2_{n_g} \\
&= \frac{n_M}{n_R} \sum_g s^2_{n_g} + n s^2_n - \left(\frac{n_M}{n_R} + \frac{n^{(2)}_M}{2}\right) \sum_g \frac{n_g}{n} s^2_{n_g} \\
&\quad - \frac{1}{n} \sum_g \tilde{n}_{M_g} s^2_{n_g}. \qquad (16)
\end{aligned}$$

Finally, we need to calculate the expectations with respect to the first phase sampling design s, but since the n_g are random, the expression cannot be further simplified. If the divisor $n(n-1)$ is included, then (16) immediately becomes

$$\begin{aligned}
\mathrm{E}(\widehat{\mathrm{V}}(\bar{y}_I)) &= \frac{n_M}{n(n-1)n_R} \mathrm{E}_s \left(\sum_g s^2_{n_g}\right) + \frac{s^2_N}{n-1} \\
&\quad - \frac{\left(\frac{n_M}{n_R} + \frac{n^{(2)}_M}{2}\right)}{n(n-1)} \mathrm{E}_s \left(\sum_g \frac{n_g}{n} s^2_{n_g}\right) \\
&\approx \frac{s^2_N}{n} - \frac{n^{(2)}_M}{2n^2} \mathrm{E}_s \left(\sum_g \frac{n_g}{n} s^2_{n_g}\right). \qquad (17)
\end{aligned}$$

It should be mentioned that we are dropping terms of order $\frac{1}{n^2}$ or smaller in our calculations.

An approximate expression for the asymptotic bias of $\widehat{\mathrm{V}}(\bar{y}_I)$ is obtained from (17) and (7) as

$$\begin{aligned}
\mathrm{B}(\widehat{\mathrm{V}}(\bar{y}_I)) &= \mathrm{E}(\widehat{\mathrm{V}}(\bar{y}_I)) - \widehat{\mathrm{V}}(\bar{y}_I) \\
&= \frac{s^2_N}{N} - \left(\frac{n_M}{n^2} + \frac{1}{n_R} - \frac{1}{n}\right) \\
&\quad \times \mathrm{E}_s \left(\sum_g \frac{n_g}{n} s^2_{n_g}\right). \qquad (18)
\end{aligned}$$

The first term will be further ignored for the present article. It is of smaller order if $\frac{n}{N} = o(1)$. We will propose an adjusted variance estimate that corrects for the bias (18). The bias correction wil be obtained by modifying an existing replication variance that produces unbiased variance estimators when there is no nonresponse.

4 Adjusted Variance Estimator

We start from a (possibly grouped) jackknife procedure. The jackknife procedure is carried out as follows. We delete a set of units from the whole set of n units. Suppose a deleted set has d units. We do not necessarily consider all possible subsets of d units out of n units in our sample. Some appropriate set is chosen, call it as $\mathcal{H}$. Thus, $\mathcal{H}$ consists of all d-tuples of indices such that the corresponding d-tuple of units was deleted in one of the jackknifed samples. Formally,

$$\begin{aligned}
\mathcal{H} &= \{(i_1, \ldots, i_d) : 1 \le i_1 < \ldots < i_d \le n \\
&\quad : (i_1, \ldots, i_d) \text{ is deleted}\}.
\end{aligned}$$

A jackknife variance estimator can then be defined as

$$\widehat{\mathrm{V}}_{JK} = \sum_{h \in \mathcal{H}} c_{[h]} \left(\bar{y}_{[h]} - \bar{y}_I\right)^2, \qquad (19)$$

where $\bar{y}_{[h]}$ is the sample mean of the hth jackknife replicate, and the $c_{[h]}$ are appropriate weights corresponding to the hth deleted set in $\mathcal{H}$. In case of delete-1 jackknife and with no nonresponse, this jackknife estimator is equivalent to the estimator $\widehat{\mathrm{V}}(\bar{y}_I)$ defined earlier and is asymptotically unbiased for $\mathrm{V}(\bar{y}_I)$. We are going to adjust this estimator to account for nonresponse and imputation.

Any missing observation that was imputed by using two randomly selected observations from the response set (i.e. those in the $M^{(2)}_g$) can be written as

$$y^{**}_{gj} = \frac{y^*_{gj1} + y^*_{gj2}}{2} : \quad j \in M^{(2)}_g.$$

Suppose we modify this imputed value by removing one of the two values and only using the remaining one. After changing the imputation weights assigned to y^*_{gj1} and y^*_{gj2} to reflect this change, the new value is

$$\tilde{y}^{**}_{gj} = y_{gj2}.$$

The difference between the original and the modified imputed values is

$$y^{**}_{gj} - \tilde{y}^{**}_{gj} = \frac{y^*_{gj1} - y^*_{gj2}}{2}.$$

This difference, averaged over a carefully selected set of imputed sample elements, can be used to measure the additional variability caused by the nonresponse and imputation mechanisms, as will now be explained.

Depending on a deleted d-set $h = (i_1, \ldots, i_d)$ removed from the original sample, we need to choose

q_g elements from the remaining elements in set $\mathrm{M}_g^{(2)}$ (for all g) on which we will modify the imputed values as described above. The numbers q_g are to be determined. So, let us denote the set of units (indices) in $\mathrm{M}_g^{(2)}$ on which the weights are changed (depending on our choice of $h \in \mathcal{H}$) as $\mathcal{E}_{h,q_g}$ and denote the union of this sets over g as

$$\mathcal{E}_{h,q} = \cup_{g=1}^{G} \mathcal{E}_{h,q_g}.$$

Define the mean of a jackknifed sample after changing weights as

$$\begin{aligned}
\bar{y}_{[h,q]}^{'} &= \frac{1}{n-d}\left[\sum_g \sum_{j\notin h, \notin \mathcal{E}_{h,q}} y_{gj}^{**} + \sum_g \sum_{j\in \mathcal{E}_{h,q_g}} \tilde{y}_{gj}^{**}\right] \\
&= \bar{y}_{[h]} + \frac{1}{2(n-d)} \sum_g \sum_{j\in \mathcal{E}_{h,q_g}} \left(y_{gj1}^{*} - y_{gj2}^{*}\right) \\
&= \bar{y}_{[h]} + f_{[h,q]}, \quad \text{(say)}
\end{aligned}$$

where $\bar{y}_{[h]} = \frac{1}{n-d}\sum_g \sum_{j\notin h} y_{gj}^{**}$. Define the modified jackknife variance estimator as

$$\widehat{\mathrm{V}}_{JK}^{'} = \sum_{h\in\mathcal{H}} c_{[h]} \left(\bar{y}_{[h,q]}^{'} - \bar{y}_I\right)^2,$$

which is written as

$$\begin{aligned}
\widehat{\mathrm{V}}_{JK}^{'} &= \sum_{h\in\mathcal{H}} c_{[h]} \left(\bar{y}_{[h]} - \bar{y}_I\right)^2 + \sum_{h\in\mathcal{H}} c_{[h]} f_{[h,q]}^2 \\
&\quad + 2\sum_{h\in\mathcal{H}} c_{[h]} f_{[h,q]} \left(\bar{y}_{[h]} - \bar{y}_I\right) \\
&= T_1 + T_2 + T_3, \quad \text{(say)} \qquad (20)
\end{aligned}$$

where $\{c_{[h]} : h \in \mathcal{H}\}$ are the originally defined weights for this jackknife estimator in (19). These weights do not depend on the choice of units which are being modified, as they only depend on the deleted units h. In the expression for $f_{[h,q]}$, the terms y_{gj1}^{*} and y_{gj2}^{*} are $i.i.d$ and hence $\mathrm{E}_{im}(T_3|R) = 0$. Since, $|\mathcal{E}_{h,q_g}| = q_g$ we have

$$\mathrm{E}_{im}(f_{[h,q]}^2|R) = \frac{1}{2(n-d)^2} \sum_g q_g s_{R_g}^2.$$

Hence,

$$\mathrm{E}_R\left(\mathrm{E}_{im}\left(T_2|R\right)|s\right) = \frac{q}{2(n-d)^2} \sum_{h\in\mathcal{H}} c_{[h]} \sum_g \frac{n_g}{n} s_{n_g}^2.$$

Write $\sum_{h\in\mathcal{H}} c_{[h]} = C_{\mathcal{H}}$. From (20), taking expectation of both sides we find

$$\begin{aligned}
\mathrm{E}\left(\widehat{\mathrm{V}}_{JK}^{'}\right) &= \mathrm{E}\left(\sum_{h\in\mathcal{H}} c_{[h]} \left(\bar{y}_{[h]} - \bar{y}_I\right)^2\right) \\
&\quad + C_{\mathcal{H}} \frac{q}{2(n-d)^2} \mathrm{E}_s\left(\sum_g \frac{n_g}{n} s_{n_g}^2\right). \qquad (21)
\end{aligned}$$

Equating the last term in (21) with the first term in (18) we find q depending on $\mathcal{H}$ as

$$q = \frac{2(n-d)^2}{C_{\mathcal{H}}}\left(\frac{n_M}{n^2} + \frac{1}{n_R} - \frac{1}{n}\right).$$

In case of delete one jackknife, we will have $d = 1$ and $c_{[h]} = \frac{n-1}{n}$, which gives $C_{\mathcal{H}} = (n-1)$, and in that case the formula for q reduces to

$$q = 2\left(\frac{n_M}{n} + \frac{n}{n_R} - 1\right).$$

Moreover, under the assumption of (1),

$$q = 2\left(\frac{1}{\alpha_1} - \alpha_1\right). \qquad (22)$$

5 Simulation Study

We conduct a small simulation study. Consider $G = 3$, with $N(\mu_g, \sigma_g^2)$ as population groups. Fix the values, $\alpha_1 = 0.6$ and $\alpha_2 = 0.2$. Also fix $(N_1, N_2, N_3) = (500, 600, 650)$ and $n = 350$. Thus the sampling fraction is 0.2. We do a Monte Carlo simulation to find the MC estimates of $\mathrm{V}(\bar{y}_I)$, $\mathrm{E}(\widehat{\mathrm{V}}(\bar{y}_I))$ and $\mathrm{E}(\widehat{\mathrm{V}}_{JK}^{'})$. The Monte Carlo simulation size is $B = 5000$. We consider two cases corresponding to low and high imputation bias.

1. Population 1: with the groups $\mu = (-10, 0, 10)$ and $\sigma^2 = (1, 1, 1)$, shown in Figure 1.
2. Population 2: with the groups $\mu = (-1, 0, 1)$ and $\sigma^2 = (1, 1, 1)$, shown in Figure 3.

The simulation is carried out by repeating the following steps B times:

1. From the set of $N = 1750$ units, draw a without replacement sample of size $n = 350$.
2. Within the sample from the gth group of size n_g, draw without replacemt samples of sizes $n_{R_g} = \alpha_1 n_g$ and $n_{M_g}^{(1)} = \alpha_2 n_g$.
3. Impute the missing values using the described procedure. Compute the jackknife estimator $\widehat{\mathrm{V}}_{JK}$ in (19), which is equivalent to $\widehat{\mathrm{V}}(\bar{y}_I)$. Also calculate the imputed mean $\bar{y}_I$.
4. Find q using (22) and compute q_g's as a multinomial sample from a population of size q and proportions $\frac{n_g}{n}$.
5. Within each delete-1 jackknife replicate created, draw a random sample of q_g elements from $n_{M_g}^{(2)}$ elements and change weights on them. Carry out this procedure for n jackknife replicates, and compute $\widehat{\mathrm{V}}_{JK}^{'}$.

μ	σ^2	$V(\bar{y}_I)$	$E(\widehat{V}(\bar{y}_I))$	$E(\widehat{V}'_{JK})$
$(-10, 0, 10)$	$(1, 1, 1)$	0.1540	0.1511	0.1535
$(-1, 0, 1)$	$(1, 1, 1)$	0.0064	0.0035	0.0059

Table 1: MC Estimates for the two different populations.

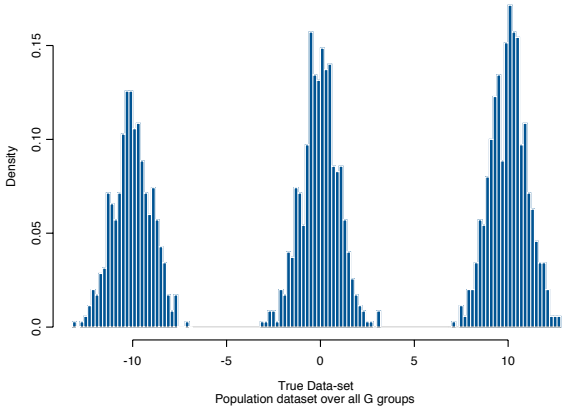

Figure 1: $\mu = (-10, 0, 10)$ and $\sigma^2 = (1, 1, 1)$: Histogram for true data of size $N = 1750$ generated from the populations.

Table 1 shows the MC estimates obtained in the simulation study. Figures (2) and (4) displays the histograms for $\widehat{V}(\bar{y}_I)$ (left) and $\widehat{V}'_{JK}$ (right) with the MC estimate of the true variance $V(\bar{y}_I)$ (shown as a vertical line) for the two populations. It can be seen that for Population 1, which has widely separated population groups, the bias due to imputation is low and the naive variance estimator performs as well as the modified jackknife estimator in estimating the true variance. But for Population 2, with largely overlapping population groups, the bias due to imputation is high and hence the modified estimator performs better than the naive one, which underestimates the true variance. This shows that the suggested imputation method performs well, both when significant bias is present and when it is not.

6 Conclusion

In this paper, we have presented a jackknife based variance estimator of the sample mean that adjusts for nonresponse and imputation, when observations are selected from a grouped population and nonrespondents are imputed using fractional imputation.

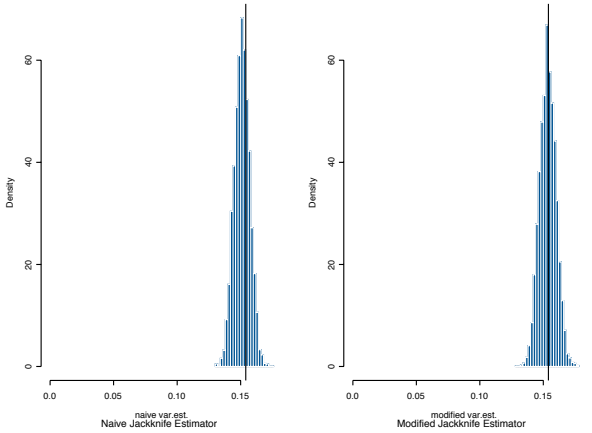

Figure 2: $\mu = (-10, 0, 10)$ and $\sigma^2 = (1, 1, 1)$: Historgram showing MC distribution of naive jackknife (left side) and modified jackknife (right side) variance estimators. The vertical line denotes the MC expected value of true variance.

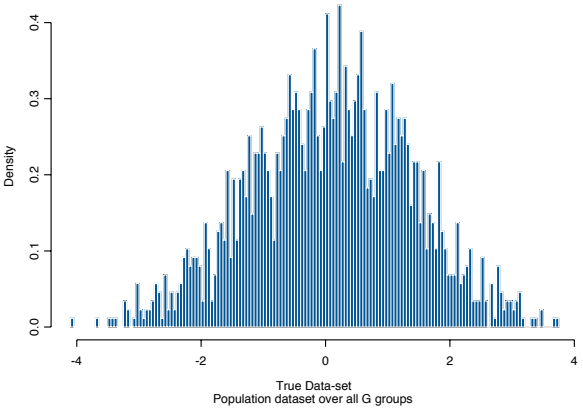

Figure 3: $\mu = (-1, 0, 1)$ and $\sigma^2 = (1, 1, 1)$: Histogram for true data of size $N = 1750$ generated from the populations.

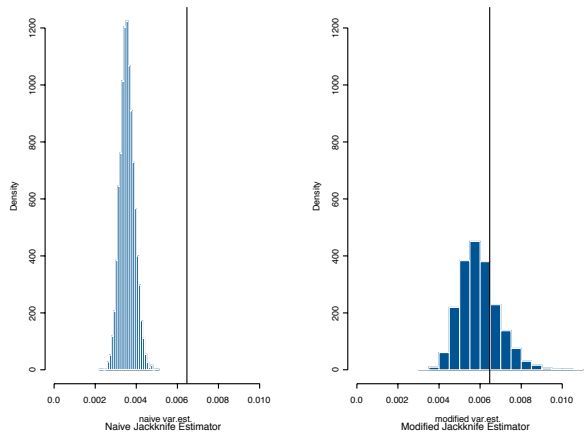

Figure 4: $\mu = (-1, 0, 1)$ and $\sigma^2 = (1, 1, 1)$: Historgram showing MC distribution of naive jackknife (left side) and modified jackknife (right side) variance estimators. The vertical line denotes the MC expected value of true variance.

The proposed method does this by adjusting the imputation weights on a certain number of imputed values in each jackknife variance replicate. Under a simple sampling and imputation scheme, we derived a formula for the number of replicates to be adjusted, and we applied this approach in a simulation experiment. The simulation results show that the method works well in different settings. Specifically, the proposed method adjusts for the imputation bias and works much better than the unadjusted variance estimator in case where there is high imputation bias (overlapping population groups). In case where there is low bias (clearly delineated population groups) and hence no need to adjust the estimator, the method works no worse than a naive variance estimator that ignores the imputation mechanism. In the future, we wish to apply the method to more complex situations where the sampling design and the imputation methods are more complex, such as the NRI.

References

Chen, J. and J. Shao (2001). Jackknife variance estimation for nearest neighbour imputation. *Journal of the American Statistical Association 96*, 260–269.

Fay, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association 91*, 490–498.

Kalton, G. and L. Kish (1984). Some efficient random imputation methods. *Communications in Statistics: Theory and Methods 13*, 1919–1939.

Kim, J. K. and W. A. Fuller (2004). Fractional hot deck imputation. *Biometrika 91*(3), 559–578.

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis with Missing Data*. New York, USA: Wiley.

Nusser, S. M. and J. J. Goebel (1997). The National Resources Inventory: A long-term multi-resource monitoring programme. *Environmental and Ecological Statistics 4*, 181–204.

Rao, J. N. K. and J. Shao (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika 79*, 811–822.

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology 18*(2), 241–252.

Tollefson, M. and W. A. Fuller (1992). Variance estimation for samples with random imputation. In *ASA Proceedings of the Section on Survey Research Methods*, pp. 758–763. American Statistical Association.